User Models and Regression Methods in Information Retrieval From the Internet

Jacek Brzezinski *

Institute for Applied Artificial Intelligence School of Computer Science,
Telecommunications and Information Systems
DePaul University, Chicago, USA

Abstract. This research focuses on learning a semantic representation of the user's information needs from a database of relevant documents and queries the in the presence of hierarchically structured semantic classes and lexical databases. The resulting user model will be enhanced by regression methods applied to capturing the syntactic structure of the documents.

1 Introduction

In traditional Information Retrieval (IR) documents are retrieved by literally comparing words from a query with words from documents. However, the natural language ambiguity makes these methods inaccurate. Internet search engines, in response to a query, often provide a long list of documents that share terms with a query but are not relevant to the users' information needs. We propose an approach that tries to overcome the problems of natural language ambiguity through the combination of evidence from four sources: queries, an external categorization hierarchy, a set of relevant documents, a lexical database such as Wordnet (see, e.g. Miller, 1990). The resulting knowledge base will constitute a user model. User models will serve as a source of semantic context information for expanding queries and filtering tasks. We will apply regression methods to model relations between the semantics and the structure of related documents. The model generation process will not require an explicit feedback from the user.

2 Problem statement

Most users submit short, abbreviated queries to search engines. As a result they receive a long list of ”hits”. The retrieved documents share terms with a query but due to the polysemy in natural language, the documents may not be related to the user‘s information needs. Also, because of the synonymy property of natural language, many relevant documents do not share terms with a query, so those documents will not be retrieved. The system we are working on is a client side agent that helps a user to retrieve relevant information from the Internet by performing three basic functions: unsupervised generation of user models for query expansion and filtering of incoming documents.

* This research has been inspired by the members of my Ph.D. committee: Dr. S. Lytinen (DePaul University), Dr. G. Knafl (DePaul University), Dr. B. Mobasher (DePaul University).

We propose an alternative for the ”bag-of-words” approach to Information Retrieval (IR) in which evidence from text documents is represented by a high dimensional vectors of weights (see, e.g. Salton, 1989). Despite the proven robustness of the vector space methods, it is often difficult to infer that two different terms are semantically close. The co-occurrence statistic methods find relationships between terms that are local to the corpora. Instead, we apply regression methods for modeling senses of terms in the presence of hierarchically structured semantic classes. The regression modeling is intended to capture the syntactic structure of the documents, whereas the external databases will be used to classify semantics of the context.

3 Overview of the system

We are interested in developing user models in an unsupervised fashion. In our system we will apply the unsupervised approach proposed by Morita and Shinoda (1984) for collecting relevant documents without an explicit relevance feedback. The collected documents will be subjected to a semantic analysis based on three sources of information: a lexical database, a classification hierarchy e.g. YAHOO (`http://www.yahoo.com`), The Library of Congress Classification System (`http://geography.miningco.com/library/`) etc., and a database containing queries submitted by a user.

We will use Wordnet as a lexical resource. Wordnet can be thought of as an extension of a thesaurus. Each lexical item belongs to one or more ”synonym sets” (synsets). Synsets are then connected by several types of links: hypernym, hyponym, meranym etc. Wordnet will be used for finding synsets for terms from the relevant documents. Classification hierarchies, included in the system, are the source of domain knowledge and will be used for semantic tagging of relevant documents. The semantic tags are sets of topics, organized in hierarchies of increasing specificity, associated with a term or a passage. The regression modeling will be used learn relations between the semantic tags and sets of terms present in the relevant documents. Those models will be utilized to automatically categorize terms or sets of terms with respect to a hierarchically structured response variable. The resulting user model is intended to be a low dimensional semantic representation of the relevant documents accompanied with a model of systematic relations between terms and structured classes. The model will be used for semantically precise query expansion and document filtering.

References

Miller G., (1990) WORDNET: An Online Lexical Database. *International Journal of Lexicography* 3 (1)

Morita M., Shinoda Y. (1984). Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. In Croft B., van Rijsbergen C. J., eds.,*Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM 272–281.

Ott L., R. (1988) *An Introduction to Statistical Methods and Data Analysis*. Duxbury Press. Fourth Edition.

Rich E. (1983) Users are individuals: individualizing user models. *International Journal of Man-Machine Studies*. 18:199–214.

Salton G. (1989) *Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Publishing Company. 313–326.